リアルタイムMRI調音動画データの閲覧および解析環境の開発

著者	浅井 拓也, 菊池 英明, 前川 喜久雄
雑誌名	言語資源活用ワークショップ発表論文集
巻	6
ページ	108-124
発行年	2021
URL	http://doi.org/10.15084/00003484

リアルタイム MRI 調音動画データの閲覧および解析環境の開発

*浅井拓也 (早稲田大学)
†菊池英明 (早稲田大学)
‡前川喜久雄 (国立国語研究所コーパス開発センター)

Development of browsing and analysis environment for real-time MRI speech articulation movie data

Takuya Asai (Waseda University)
Hideaki Kikuchi (Waseda University)
Kikuo Maekawa (National Institute for Japanese Language and Linguistics)

要旨

近年, 音声の調音運動を MRI 装置を用いて, リアルタイムに撮像することが可能になった. リアルタイム MRI (以下 rtMRI) データは, 声道の正中矢状面全体の情報が含まれ, 調音音声学の再構築を促す可能性を秘めている. しかし, 収集されたデータに対する転記方法や分析方法は整備されているとは言いづらく, 単にデータを公開するだけでは rtMRI データに基づくた調音音声学研究は普及しにくいと予想される. そのため, 我々は過去に rtMRI データ閲覧ツールとして MRI Vuewer の設計と開発を行った. 上記ツールは rtMRI データの音声的, 時間的側面の転記機能を有するものであったが, 画像的, 空間的側面の転記機能に不足があった. 本研究では近年行われた rtMRI データを用いた研究から画像的, 空間的側面の転記に必要な機能を整理し, rtMRI データ解析ツールとして MRI Vuewer の再設計と再実装した結果を報告する.

1. 背景

調音データは, 人間の音声生成メカニズムの解明や, その生体力学的特性の解明, そして言語的基盤に関する重要な情報源である. 一方で, 現実的な速度, 情報量をもつ有用な調音データの取得は長年の課題である. 調音運動の測定に使用されてきた手法には, X 線マイクロビーム (Kiritani et al. 1975) やエレクトロパラトグラフィー (Hardcastle 1972), EMA (Perkell et al. 1992, Wrench 2000), そして, 超音波断層撮像法 (Stone and Davis 1995, Whalen et al. 2005) などが存在する. これらの手法は毎秒 100 サンプル以上のサンプリングレートで, 調音運動を記録可能であるが, 侵襲的であったり, 数個のセンサの位置情報だけを提供するものであったりし, 音声発話時の声道の正中矢状面全体を安全かつ高速に計測することは困難な課題であった. それに対し, MRI 装置を使用した計測では, 喉頭・咽頭を含めた声道の正中矢状面全体の情報が含まれる. 特に近年では, MRI 装置の性能向上および高度なサンプリング技術

* qh73xe@ruri.waseda.jp
† kikuchi@waseda.jp
‡ kikuo@ninjal.ac.jp

の適用によって, リアルタイムでの MRI 動画撮像が可能になってきている (Ramanarayanan et al. 2013a). 日本では ATR Promotions の脳活動イメージングセンタが, 正中矢状断面に限定した動画を毎秒 14 ないし 28 フレームで撮像するサービスを提供している. 我々は 2017 年度から JSPS 科研費の補助をうけ, 日本語音声の rtMRI 動画データベースを構築しており, そのデータの一部が一般公開される.

しかし, rtMRI 動画データベースは, 単にデータを公開するだけでは rtMRI 動画に基づく調音音声学研究は普及しにくいと予想される. 具体的には MR 画像の標準フォーマットである DCM は画像専用であるため, 実験時に別途収録した音声信号をダビングし, 動画化する必要がある. このプロセス自体が多くの研究者にとって高いハードルになることが挙げられる. また, 作成された動画を検索し, 検索された画像に対して分析用情報を転記するには, 特殊化されたソフトウェアが必要になることも問題である.

そのため, 上記科研費研究では rtMRI データの解析環境の整備も進めている. 本稿では調音運動計測データに対するアノテーションの環境を提供を目標にした, 動画データ閲覧転記ツールである MRI Vuewer[(1)] の設計と実装を報告する.

2. rtMRI データの解析手法

本章では, アプリケーション設計の前提として, 調音運動解析にはどのようなデータ操作が必要になるかを整理する. Ramanarayanan et al. (2018) は近年の rtMRI 研究を整理し, 調音データの分析手法を, 処理方法の実装の難しさ, および出力表現の抽象度によって (i) 基底分解または行列因数分解に基づく解析, (ii) 関心領域 (ROI) に基づく解析, (iii) グリッドに基づく解析 (iv) 輪郭に基づく解析 に分類した.

最初の手法は, 元の画像全体を操作し, 比較的抽象的な時空間基底関数を取得する手法で, 主成分分析や畳み込み非負行列因子分解などを含む (Ramanarayanan et al. 2013b, 2016). この手法を実施するには関心のある時間的イベント (たとえばある音素の発話) を記述し, その時刻の画像データを取得する必要がある.

ROI ベースの手法では, 特定の関心領域を手動で定義する必要がある. この場合, 手動で設定されたピクセル座標の変動や, 共分散を観察し, 関心のある様々な言語的, 臨床的問題に対する洞察を得たり, さらに詳細にモデリングするための中間関数を提供する (Lammert et al. 2010, Tilsen et al. 2016). この分析手法は実装と解釈が比較的容易ではあるが, 使用される ROI の妥当性に関して, 解析者の解剖学的構造に対する専門知識が大きく影響することに注意が必要である. また, rtMRI データは発話者自身の個人性や発話時の姿勢により, 記録される画像のおける ROI の位置が変化するため, 適切な正規化処理が必須になる. そのため, この解析を実施するためには, 特定の関心のある時間的イベントのほかに, そのイベント内部における ROI の位置および, 正規化を施すための何らかの参照点を保存, 出力する必要がある.

3 番目の方法は, 適切な参照線系を画像に重ね合わせる手法である. この参照線と軟組織の交点を計算することで声道断面積関数を抽出する (Maeda 1979, Proctor et al. 2010). この

(1) Vuewer の表記はアプリケーション作成時に使用したライブラリに由来する.

方法は, 主に音響学の中で発展してきた声道断面積関数を近似する指標を取得可能であるため, 解析結果と音響特徴量との関連を観察するのに役立つ手法であるが, 一般には 3 次元で撮像された調音データを前提にすることが多く, 2 次元画像の MRI から, 参照線を引くことは困難な課題である. そのため, この解析手法に関しては本アプリケーションの利用想定から除外した.

最後の分析方法は, 音声生成に関与する組織境界の抽出を行う方法である. このような方法の例には, 画像セグメンテーション (調音器官の輪郭を生成する) や, 声道の収縮変数や幾何学的な調音座標など, 輪郭から導出できる高次特徴が含まれ, 解剖学的構造のより詳細な仕様を分析に直接使用したり, モデリングの中間関数として使用することが可能である (Sampaio and Jackowski 2017, Raeesy et al. 2013). 一般に組織境界の推定には, 顔器官検出やセマンティックセグメンテーションなど, 教師情報を利用した機械学習手法が用いられることが多い (たとえば (Takemoto et al. 2019)) が, 教師なしおよび半教師あり声道画像セグメンテーションに関する研究も増えている (Raeesy et al. 2013).

上で概括した基本的な解析手法は, いずれの手法を用いる場合においても, 解析者は複数の動画データに対し, 関心のある音声的, 時間的な区間を決定する必要がある. また, その関心をある区間において一枚ないし, 複数枚の画像を選択する. 加えて ROI に基づく解析以降の解析においては選択された画像を空間的に分離し, その座標や輝度値を解析するまた輪郭に基づく解析の場合には, 分割された画像および元画像を使用し, 輪郭推定モデルの構築を行う. モデルの構築方法や指定された ROI の解析方法, 画像の正規化方法等は, 解析者の興味に応じて決定されるものである. そのため, 本アプリケーションの機能要件からは除外し, 動画の時間的な区間および画像の決定, 画像の空間的な境界情報の管理, そして, それらの転記情報を前提としたファイル操作を行えるようにすることが, 本アプリケーションの目的である.

3. MRI Vuewer

3.1 データ設計

図 1 に本アプリケーションで管理を行うデータ構成を示した. 本アプリケーションは基本的には 1 つの動画を操作しながら, 大きく 2 種類の転記情報を記述する. 1 つは時間的情報 (時系列アノテーション) であり, もう 1 つは空間的情報 (フレームアノテーション) である.

時間的情報は, Ramanarayanan et al. (2018) が分類した 4 種類すべての方法において使用される情報であり, 解析者の興味により, さまざまな単位を取りうる (たとえば単語境界やモーラ境界, 音素境界等). そのため, 1 つの動画に対し複数の時間的単位を管理する必要がある. 複数の時間的単位を管理する方法として, 特に音声研究の中でよく使用されるツールには ELAN (Wittenburg et al. 2006) や Praat (Boersma and Weenink 2018) が存在するが, これらのツールにおいては Tier (層) の概念が用いられる. 本アプリケーションもこれらのツールを踏襲し, 時系列アノテーションテーブルでは “tier_name” 情報を管理する. ここで, 本アプリケーションが管理する時系列アノテーションが, CSJ-RDB (Koiso et al. 2014) のように層情報を明示的に分離せず, 1 つの大きなテーブルとして表現されていることに注意されたい. これは本アプリケーションの特性上, 解析者の興味に合わせて層の設定を行う必要があるため,

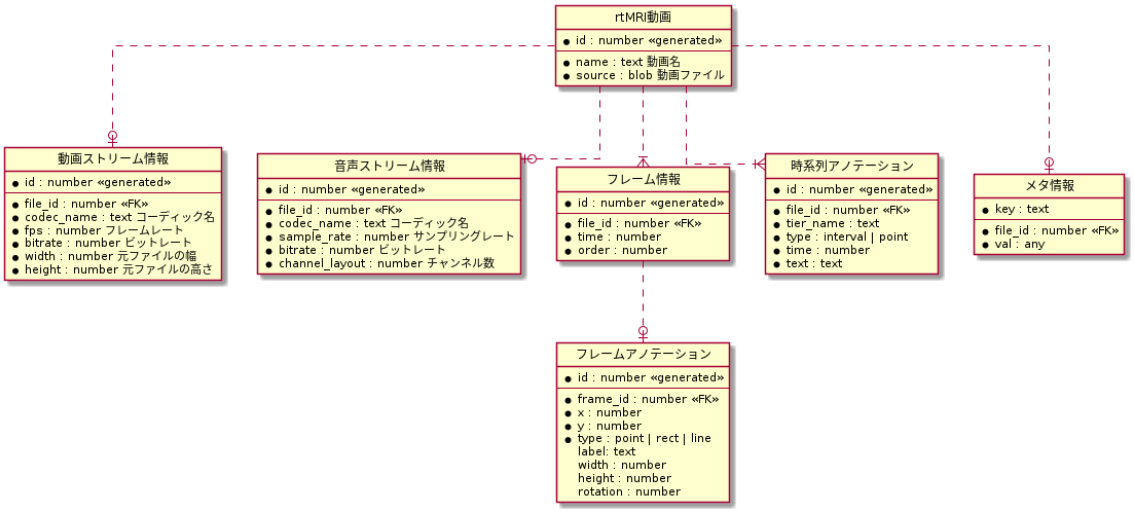


図 1　MRIVuewer における実体関連図．時系列アノテーションは動画そのものに対し，リレーションが存在することに対し，フレームアノテーションは，動画そのものと直接的なリレーション関係にないことに注意．また，両アノテーション情報ともにそれぞれ単一のテーブルとすることで，データ追加および検索を簡便に実施可能となるように設計されている．

事前に層そのものを定義できないためである．また，1 つの大きなテーブルとして表現することでファイルを横断した検索やデータ出力を行いやすくするという狙いもある．

時間的情報に関しては，特に調音データであることを考えると，調音器官同士の接触のような高々 1 フレームで終了するイベントと，顎の開閉のようにある程度の長さを持つイベントに大分される．これらの時間的情報は音響情報を参照して決定されることもあるが，画像情報や，その差分情報を元に決定されることもある．ここで Praat を参考にすると，開始終了時刻を管理する interval tier と一点の情報のみを管理する point tier のようにテーブルレベルで概念を分けて表現している．一方で，本アプリケーションでは，時系列アノテーションという大きなテーブルの中に属性 “type” を有し，この属性を利用し，アプリケーションの表現を変更した．

また，長さをもつイベントに関しては，一般にアナログな値をとることが多いが，そこに含まれる画像は有限個であることに注意が必要である．最終的に解析を行う対象は画像であることを考えると，時間的情報も動画のフレームレートに合わせた記述を行う必要がある．ここに関しては，データ構造上の関連を持たず，アノテーション時の操作方法として実装を行った．

解析者は，各画像単位において空間的な境界を決定する．これは特定フレームの画像に対する転記であるため，フレームアノテーションと名付け，データベースを設計した．ここで記録される空間的情報は，ROI 解析のように直接解析されることもあれば，輪郭推定ための教師情報として利用される場合もある．また，輪郭情報を入力とし，動画と重ね合わせた後に ROI 的な解析を実施が実施されることも想定される．言い換えれば時間的情報，空間的情報の両単位で，画像ファイルを柔軟に出し入れできることが望まれる．そのため，フレームアノテーションは動画に直接関連を持つものではなく，フレーム情報と関連するものと定義した．

この空間的情報は，ある種の矩形として表現することも可能であるし，特定の 1 つの座標と

して点で表現する場合もある．加えてデータを質的な側面でとらえる場合, 調音器官の境界など解析に直接的に使用する情報と, 正規化処理を実施するための参照点 (ないし矩形, または線) として使用する情報も存在する．これらの幾何学的な特性は, オプショナルな属性として記録され, 属性 “type” によって使い分けられるように設計されている．

ここまでは単一動画に対する機能要件を列挙したが, 実際の解析においては, 単一動画のみを解析の対象とすることは考えにくい．そのため, 複数動画を跨いで, 時間的イベント情報および空間的境界情報を検索したり, その検索結果に応じて記述を追加, 更新したり, 各種記述された情報を出力する機能も必要もある．これに関して, 本アプリケーションでは時系列アノテーションおよびフレームアノテーションを大きな単一のテーブルとして表現を行うことで横断的な検索を可能とした．また, 動画に対する非構造的な情報 (たとえば発話者 ID 等) を管理するためのテーブルとしてメタ情報テーブルを作成した．

3.1.1 モジュール構成

本アプリケーションを機能的にみると, データ保存モジュール, I/O モジュール, 音声解析モジュール, 画像解析モジュール, 動画再生モジュールによって構成される．

データ保存モジュールは前節で説明したデータ構成をとるデータベースへのデータ登録および, 検索を担う．なお, Web アプリケーションは一般的に, サーバ, クライアント間での通信が必要不可欠なものであるが, 本アプリケーションにおいては, Indexed Database (W3C 2021) と呼ばれる各ブラウザに固有のデータベースを使用している．

I/O モジュールはファイルのフォーマット変換を行うモジュールである．たとえば動画データを取り込む際に, その動画に独自の情報 (FPS やもともとの画像サイズ等) を解析したり, 動画データを画像データに変換したりする．また, データベースに登録されたアノテーション情報のフォーマットを変換し, XLSX ファイルや TextGrid ファイルとして出力する．

音声解析モジュールは動画ファイルを受け取り, そこに記録されている音響データの解析を行う．rtMRI 動画を解析する場合, 特に音声スペクトルとそれを生成した調音データを同期的に観察することが多い．そのため, 音声解析モジュールでは音声スペクトログラムの出力を可能とする．また, rtMRI 動画には撮動音が比較的大きなノイズとして記録される．そのため, 簡易なノイズ除去機構を実装する．なお, アプリケーション作成の効率化のため, 音声スペクトルの算出処理に関しては wavesufer.js (Guisch and thijstriemstra 2018) を利用した．また, ノイズ除去処理に関しては動画, 音声の加工を行うためのフリーソフトウェアである FFMPEG (Tomar 2006) を WebAssembly (MDN web docs 2018a) の技術を利用し, ブラウザ上で実行可能なバイナリコードに変換し利用した．

正規化に用いられる参照点は一般に硬組織に対し付与される．硬組織は何らかの解剖学的なランドマークを持ちうるため, 簡易な画像加工を行うことにより, 座標を決定しやすくなる場合がある．そのため, 画像に対するフィルタ処理も行えることが望ましい．画像解析モジュールはこのフィルタ処理を実施する．画像データは一般に縦横 2 次元の配列に対し, それぞれ 3 次元の色情報が記録されている．この配列演算の実施時間を高速にするため, オープンソースのコンピュータビジョン向けライブラリである OpenCV(Bradski 2000) を利用した．

動画再生モジュールは, 動画の再生, 停止, フレーム単位での移動処理を担うモジュールであ

る．Web アプリケーションの場合動画の再生は HTML Video 要素 (MDN web docs 2018b) を介して実行されるが, 本アプリケーションの場合, 特にフレームアノテーション結果を動画再生に合わせて表示を行いたいという需要が存在する．そのため, 本アプリケーションにおいては HTML Video 要素上にその時刻のフレームアノテーションのみを描く透明な要素を生成し, これをパラパラマンガのように切り替えることで上記の需要に応えている．

3.2 実装および利用例

本節では想定される利用シーンに準拠し, 本アプリケーションの実装を説明していく．

基底分解または行列因数分解に基づく解析を実施する場合, 解析者は 1 つ以上の動画を本アプリケーションに登録する．その後, 解析者の興味のある時間的イベントに従い, 登録したすべての動画に対し, 時系列アノテーションを施す．最後に動画に対し横断的に検索を掛け, 興味のある時間的イベントが発生している時刻に対応する画像データを取得する．

関心領域 (ROI) に基づく解析を行う場合に関しても, 時系列アノテーションを施すまでの処理は同様である．ただし, その後, 時間的イベントが発生した時刻に対応するフレームの画像を一枚, ないし複数枚選択し, それらの画像に対し ROI や 正規化のための参照点を空間的情報として記述する．最後に動画に対し横断的に検索を行い, 興味のある時間的イベントが発生している時刻に対応する空間的情報を取得する．

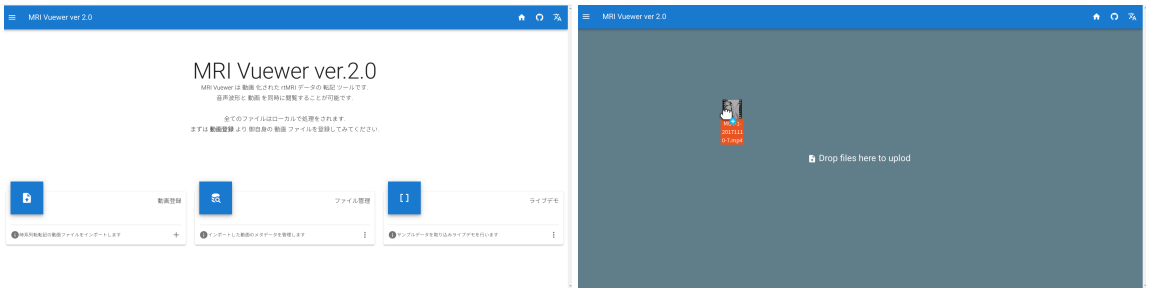
輪郭に基づく解析を行う場合, 本アプリケーションの利用シーンは, 輪郭推定機を作成するための教師情報を生成する場合と, 推定された輪郭情報を動画に反映し, 確認する場合が考えられる．前者の場合, 解析者は関心領域 (ROI) に基づく解析を rtMRI に利用する場合と同様の処理を行う．ただし, 最後に取得するファイルは空間的情報のみでなく, その空間的情報が付与されたフレームに対応する画像ファイルも必要である場合が多い．後者の場合には, 解析者は, 作成された境界を空間的情報として事後的に登録できる必要がある．

まとめると, 本アプリケーションの利用シーンは, 動画登録, 時系列アノテーション, フレームアノテーション, 動画を横断する検索, ファイル出力のいずれかである．これらの要求は, 本アプリケーションにおいて, ホーム画面, 時系列アノテーション画面フレームアノテーションダイアログ, データマネージャー画面によって実現されている．

3.2.1 ホーム画面

まず, 解析者は, ブラウザ上で MRI Vuewer にアクセスする (関連 URL MRI Vuewer ver2 を参照)．アクセスに成功するとホーム画面が表示される (図 2a)．最も簡単な動画登録の方法は, 登録を行う動画ファイルを 1 つ選択し, 開始画面上にドラックすることである．動画をドラックすると図 2b の画面に変わる．その状態でドロップをすると動画の登録が実施される．なお, ドラック時には複数の動画ファイルをまとめて登録することも可能であるが, 処理時間や登録可能な動画数は解析者の使用する PC に依存するため, 多用は控えるべきである．ファイルの登録に成功するとサイドバー (開始画面左上にあるメニューボタンをクリックすると表示される) に, 登録された動画が表示され, 時系列アノテーション画面 (図 3) に遷移する．

ここで, 登録処理をより細かく制御することも可能である．ホーム画面左の動画登録カードにあるプラスボタンをクリックすると, 動画登録ダイアログが表示される．このダイアログは



(a) 開始画面　　**(b)** 動画ファイルドラック時

図 2　MRI Vuewer 動画登録画面. 解析者はこの画面に対し動画ファイルをドラッグアンドドロップすることでデータベースに動画情報を登録できる.

動画の選択, FPS 等の動画コーデックの確認および編集, 動画に対するメタデータ登録の 3 つの処理が実施される. この内, 動画に対するメタデータの登録はドラックアンドドロップでファイル登録を実施する場合, 実行されないことに注意されたい.

3.2.2　時系列アノテーション画面

解析者は時系列アノテーション画面 (図 3) にて, 自身の興味に応じた時間的イベントが発生した時間情報を記述する. この画面は, 動画再生コンポーネント (図 3 左上), アノテーション一覧表示コンポーネント (図 3 右上), 音声スペクトル表示コンポーネント (図 3 下), アプリケーションボタン (図 3 左下) によって構成される.

動画再生コンポーネントには, 連続する 3 フレームの MR 画像および, 動画操作用メニューが表示されている. なお, この MR 画像をクリックし, CTRL-c を入力すると選択された MR 画像がクリップボードにコピーされる.

動画操作用メニュー中央にある再生ボタンをクリックすると動画が再生あるいは停止する. 左右端にある進む, 戻るボタンをクリックすると, MR 画像は 5 フレーム分移動する. また, その内側にある進む, 戻るボタンをクリックすると, MR 画像は 1 フレーム分移動する. 拡大縮小ボタンはスペクトル表示領域を拡大したり, 縮小したりする際に利用する. アップロード, ダウンロードボタンは登録された転記情報をファイルとして保存したり, 既存のアノテーション情報を上書きする際に使用する. なお, 単一ファイルに於ける既存アノテーション情報のダウンロード可能なファイル形式には TextGrid 形式, XLSX 形式, JSON 形式が用意されている. 特に XLSX JSON 形式は時系列アノテーション情報のみならず, フレームアノテーション情報, 動画情報 (JSON 形式のみ) も構造化されダウンロード可能である.

時系列アノテーションを実施するには 1 つ以上の転記用の層を設定する必要がある. 転記層の登録はアプリケーションボタンをクリックし, その後表示されるプラスボタンを選択する. 新規時系列転記層記入欄と書かれたダイアログが表示されるので, 転記層の識別名および転記層の種類を入力する. ここで転記層には境界転記層 (開始, 終了を持つ転記情報), および, イベント転記層 (一点の時刻情報のみを持つ転記情報) の二種類が存在する.

新規時系列転記層記入欄には上記 2 つの入力項目のほかに, 時刻情報をコピーすると記述さ

れたチェックボックスが存在する．このチェックボックスは複数層を作成していく際に利用する項目である．特に音声学における時間的情報には, たとえば単語境界に対するモーラ境界のように或る種の包含関係をもつ単位が多い．そのため, 事前に時系列アノテーションの時刻情報はコピーする形で, 新規転記層を設定することを可能にしている．

転記層の作成に成功すると, アノテーション一覧表示コンポーネント上部にあるタブが増え, 音声スペクトル表示コンポーネントに階層が 1 つ増える．音声スペクトル表示コンポーネントの任意の階層をクリックすると色がオレンジに変化し, その階層に対し時系列アノテーションを行うことが可能になる．解析者の興味のある時間的イベントが発生した時刻を音声スペクトログラムや MR 画像より認定し, 該当時刻を音声スペクトログラムから選択する．その後ダブルクリックを行うと該当時刻に時間的情報を記録する．時刻情報が登録された後に該当の境界をクリックすると, クリックされた時刻にもっとも近い箇所がオレンジ色に変化する．その後, 層上部にある入力欄に任意の文字を入力し, ENTER を入力すると, 選択箇所に文字が入力される．なお, 編集層がオレンジになっている状態で TAB を入力すると, その箇所のみ動画を再生する．解析者はこれを繰り返すことで時系列アノテーションを行う．

ここで動画ファイルに含まれる音声データと解析を行たい画像データのサンプリング速度は一般に一致しないことに注意されたい．つまり, マウス操作を前提に時系列アノテーションを行うと, 記録される時刻は必ずしもフレームと一致しないものに成り得る．そのため本アプリケーションではキーボード操作による時系列アノテーションも可能としている．たとえば編集を行いたい層を選択後 j を入力すると, 動画が一フレーム分前に戻る．また k を入力すると動画が一フレーム分進む．この操作により任意時刻に移動した後に space を入力することでその時刻の情報が記録される．また CTRL-SHIFT-k または CTRL-SHIFT-j を入力すると選択されている時刻情報を前後させることが可能である．その後, i を入力すると層上部にある入力欄が活性化し, 選択ヵ所にラベルを付与できる．

時系列アノテーションを実施時に参照可能なデータは, 音声スペクトログラムおよび連続する 3 フレーム分の MR 画像であるが, たとえば舌がもっとも大きく動く時刻を境界としたい場合など, 前後フレームの差分画像を参照したくなる場合がある．MRI Vuewer ではこの要求を満たすためにフレーム間差分表示ウィンドウを実装している (図 4 前景右)．このウィンドウは時系列アノテーション画面上で左クリックを行うと表示される VUWER メニューの内 (図 4 前景左), 補助ウィンドウの項を選択し, フレーム間差分項をクリックすると表示される．

3.2.3 フレームアノテーションダイアログ

フレームアノテーションを実施するためには転記を行うフレームを選択する必要がある．この選択は時系列アノテーション画面によって実施され, 選択方法は時系列アノテーションにおいて時刻を指定する方法と同一である．特定時刻を選択したうえでアプリケーションボタンをクリックする．その際に表示されるボタン群の中央にあるフレームアノテーション開始ボタンをクリックすると図 5a のようなダイアログが立ち上がる．

解析者はこのダイアログを利用しフレームアノテーションを実施する．このダイアログは画像操作メニュー (図 5 の各パネル上部), フレームアノテーションコンポーネント (図 5b 中央部), フレームアノテーション一覧 (図 5b 下部) によって構成される．

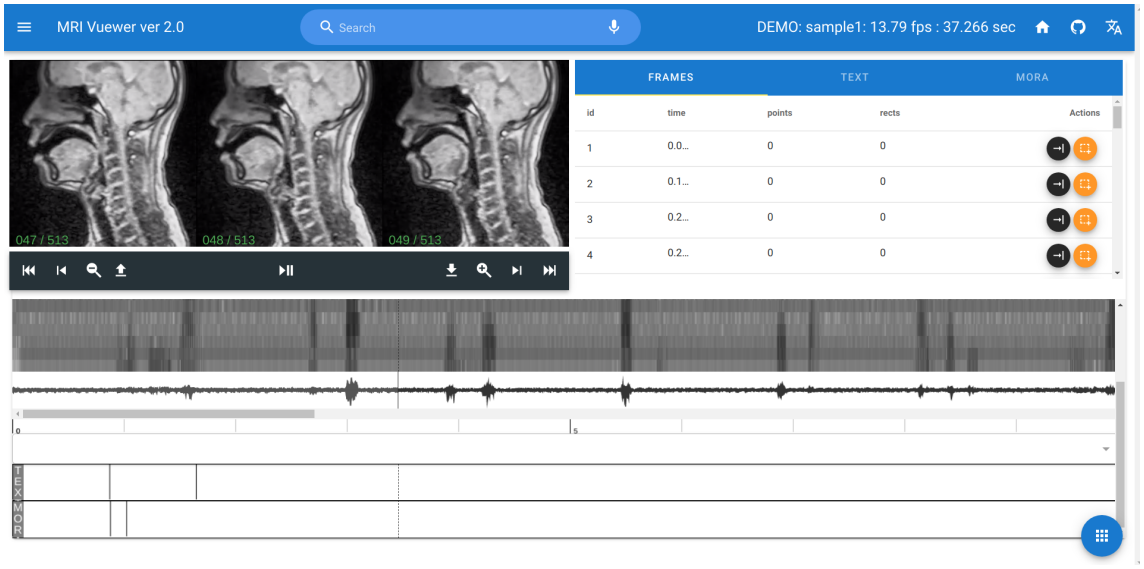


図 3　時系列アノテーション画面．この画面は連続する 3 フレームの MR 画像を表示する動画再生コンポーネント (左上), 時系列アノテーション情報を管理するアノテーション一覧表示コンポーネント (右上), 音声スペクトルを確認し, 時系列アノテーションを実施する音声スペクトル表示コンポーネント (下) によって構成される．

画像操作メニューは二段構成になっている．一段目右はフレームアノテーションコンポーネントに表示されている MR 画像を縮小拡大したり, MR 画像の指定時刻を変更する．一段目左はフレームアノテーションを行う際のモードを選択するメニューである．ここでモードには点群記述モード (図 5a), 矩形記述モード (図 5b), 補助線記述モード (図 5c), 転記削除モード (図 5d) が存在する．

二段目は MR 画像に対し簡易なフィルタ処理を施すことが可能である．上記フィルタには二値化 (図 6b), 適応的閾値 (図 6c), キャニー (図 6d), バイラテラル (図 6e), ラプラシアン (図 6f) が実装されている．これらのフィルタは, 画像処理において輪郭検出や領域検出に利用される代表的なフィルタのうち, 比較的少数のパラメータで実装可能なものを利用した．

フレームアノテーションコンポーネントは現在のモードに応じた転記を行うためのコンポーネントである．たとえば矩形記述モードの場合画像上の任意の座標をダブルクリックすることで, その箇所に点を記述できる．この点の座標情報は登録された動画ファイルの幅, 高さに変換され, データベースに登録される．また, 矩形記述モードの場合には, 任意の座標をダブルクリックすると矩形が登録される．この矩形を選択するとマウス操作にて, 高さや幅, 傾きを変更できる．補助線記述モードは最終的には点を記述するためのモードである．ただし, 最終的に登録したい点を任意の二点からなる直線上に記録する．このモードの際に像上の任意の座標をダブルクリックすると補助線を記述するための一点が決定される．異なる座標をダブルクリックすると, その座標と先の一点を結む直線が表示される．この直線上にマウスオーバーを行うと補助線は活性化 (黄色に変化) する．補助線が活性化している状態において特定の座標をダブルクリックすると, その座標がデータベースに登録される．転記削除モードは上記点, および矩

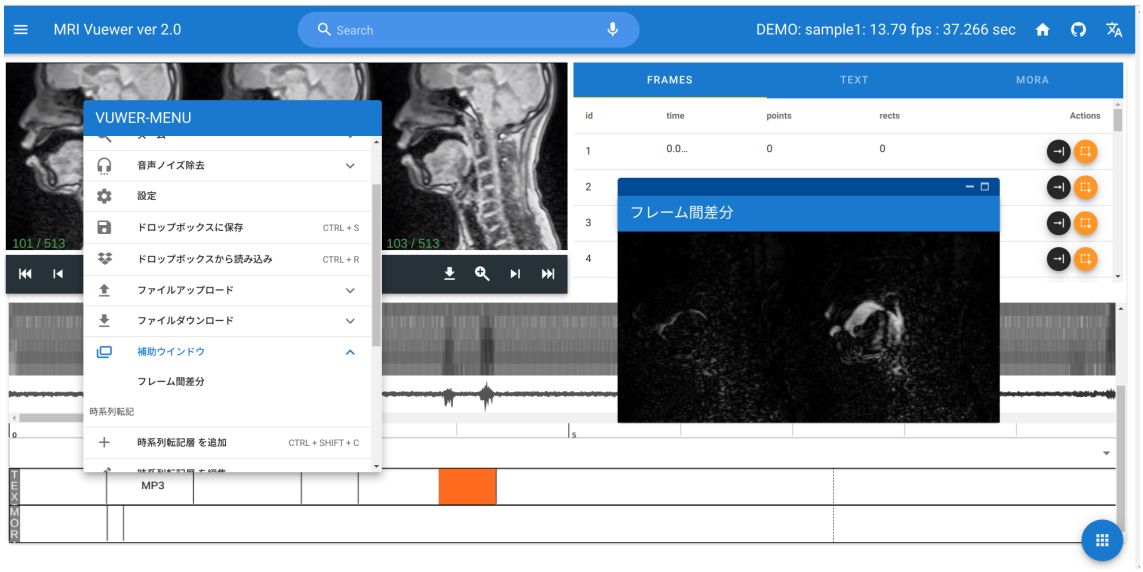


図 4　フレーム間差分ウィンドウ (前景右) および VUWER メニュー (前景左). フレーム間差分ウィンドウには動画再生コンポーネントに表示される MR 画像の, 左と中央の画像との差分画像 (前方差分画像) と, 中央と左の画像との差分画像 (後方差分画像) が表示される.

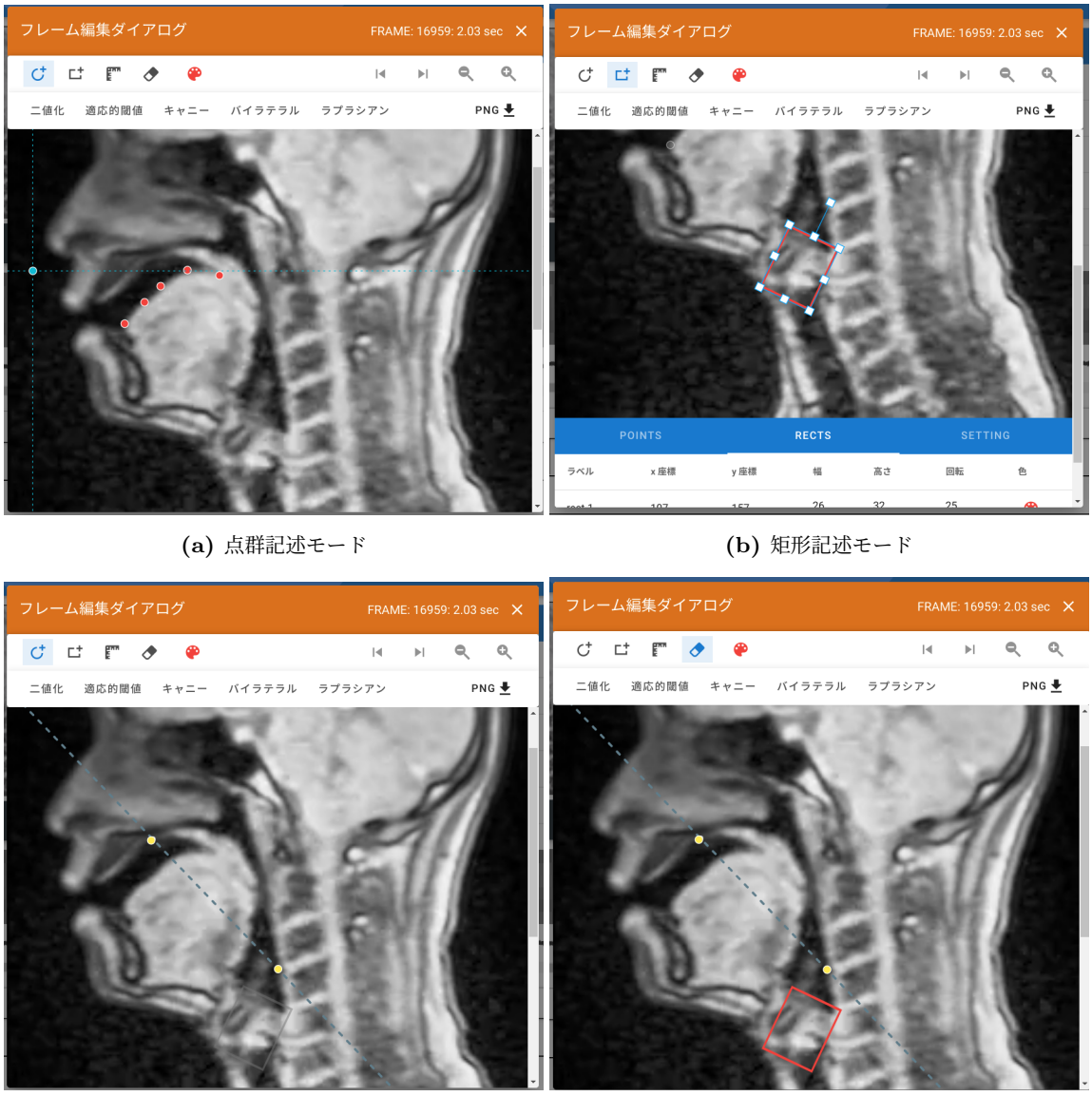
形を削除するためのモードである. このモードの際に任意の点, または矩形にマウスオーバーを行うと, 該当の点, または矩形が拡大される. この状態でダブルクリックが行われると, 該当の該当の点, または矩形が削除される.

点, または矩形には色およびラベルの登録を行うことが可能である. 新規登録を行う点, または矩形の色を変更するには, 画像操作メニュー一段目左部分の左端にあるパレットボタンをクリックし色の選択を行う. 既存の点, または矩形の色あるいはラベルの変更にはフレームアノテーション一覧を利用する. フレームアノテーション一覧上部には, POINTS, RECTS, SETTING と書かれたタブが存在する. このタブの内, POINTS を選択すると点の, RECTS を選択すると矩形の一覧が表示される. これらの一覧の内変更を行いたい点, または矩形の行のパレットボタンをクリックすると, 配色が変更でき, ラベルの項目をクリックするとラベル情報の変更が可能になる.

3.2.4　データマネージャー画面

時系列アノテーション画面およびフレームアノテーションダイアログを利用することで, 単一動画に対し, 時間的, 空間的アノテーションを行うことができる. 一方で, ここで記述したアノテーション情報を利用することを考えると, たとえば上記作業によって登録された転記情報を, 横断的に検索するなど, 複数動画に対する一括処理を行いたくなる.

データマネージャー画面 (図 7) は上記のように複数動画に対する一括処理のために作成された画面である. この画面へはサイドバーより, PAGES の項にあるファイル管理をクリックすることで遷移できる (あるいは `https://kikuchiken-waseda.github.io/mri-vuewer.ver2/mata` にブラウザアクセスしてもよい). この画面は FILE, 境界転記層, イベント転記層と書かれたタブがあり, それらの情報がテーブル形式で表現されている.

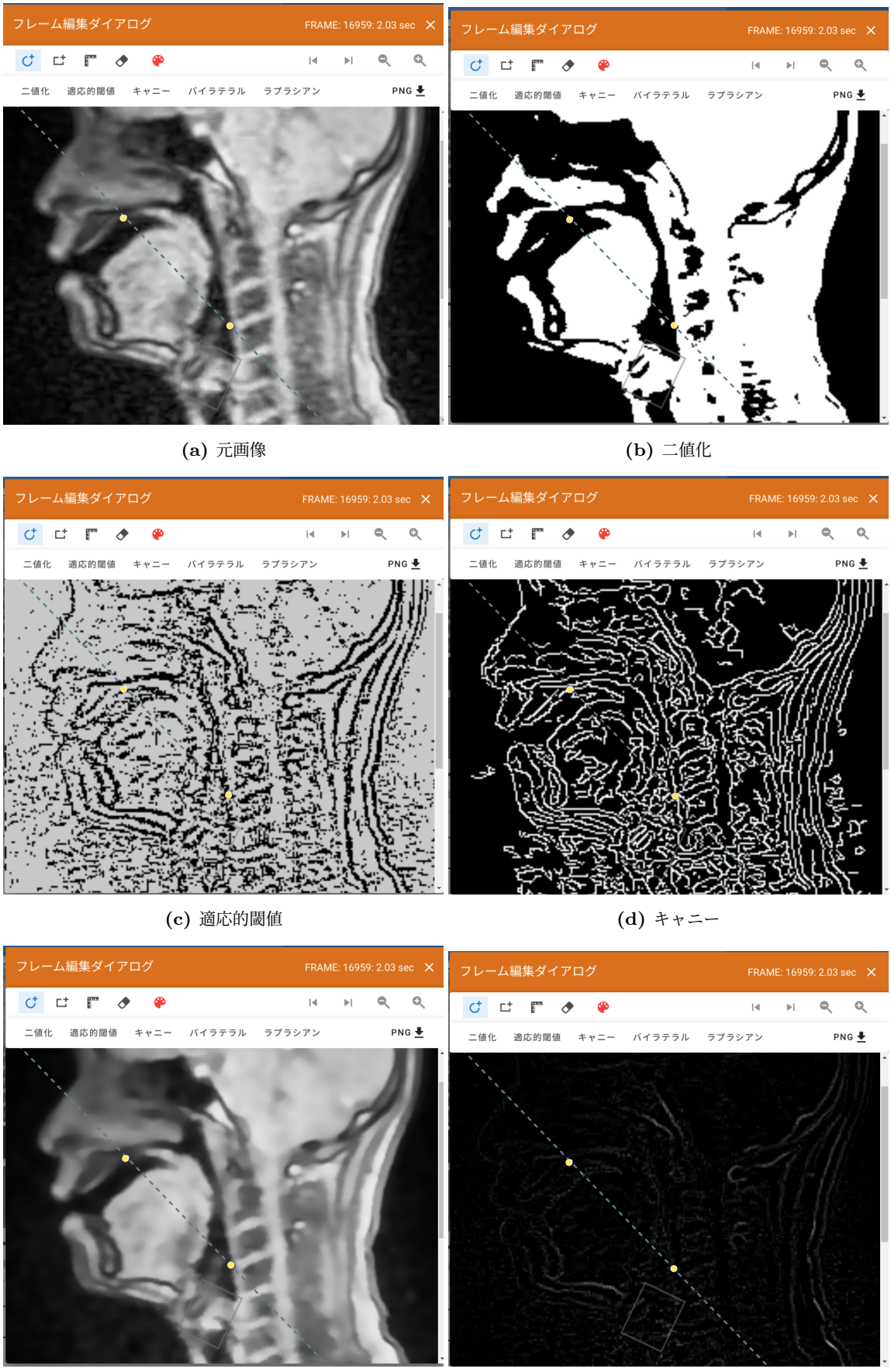


(a) 点群記述モード　(b) 矩形記述モード

(c) 補助線記述モード　(d) 転記削除モード

図 5　フレームアノテーションダイアログ．解析者は，点群記述，矩形記述，補助線記述，転記削除のいずれかのモード選択し，転記を行う．それぞれのモードにおいて操作できない空間的情報は，灰色に表現される．

FILE タブが活性化している場合に表示されるテーブルは登録された動画情報および，そのメタ情報が表示される (図 7)．ここでメタ情報はデータ設計上必ずしもすべての動画に対し，統一的な情報が保持される訳ではないことに注意されたい．そのため，このテーブルにおいては，どれか 1 つの動画に登録されたメタ情報がある場合，そのメタ情報をカラムとして追加される様に表現をしている．動画データの検索はアプリケーションバー (画面最上部にある青いバー) にある検索欄 (中央) より実施する．ここに任意の文字を入力するとメタ情報や時系列アノテーションを検索し，どこかに前方一致するファイルのみが表示される．文字列にスペース



(a) 元画像　(b) 二値化

(c) 適応的閾値　(d) キャニー

(e) バイラテラル　(f) ラプラシアン

図 6　画像フィルタ実施例．現状では輪郭検出や領域検出に利用される代表的なフィルタのうち，比較的少数のパラメータで実装可能なものを利用しているが，今後パラメータ操作を含む画像フィルタも実装予定である．

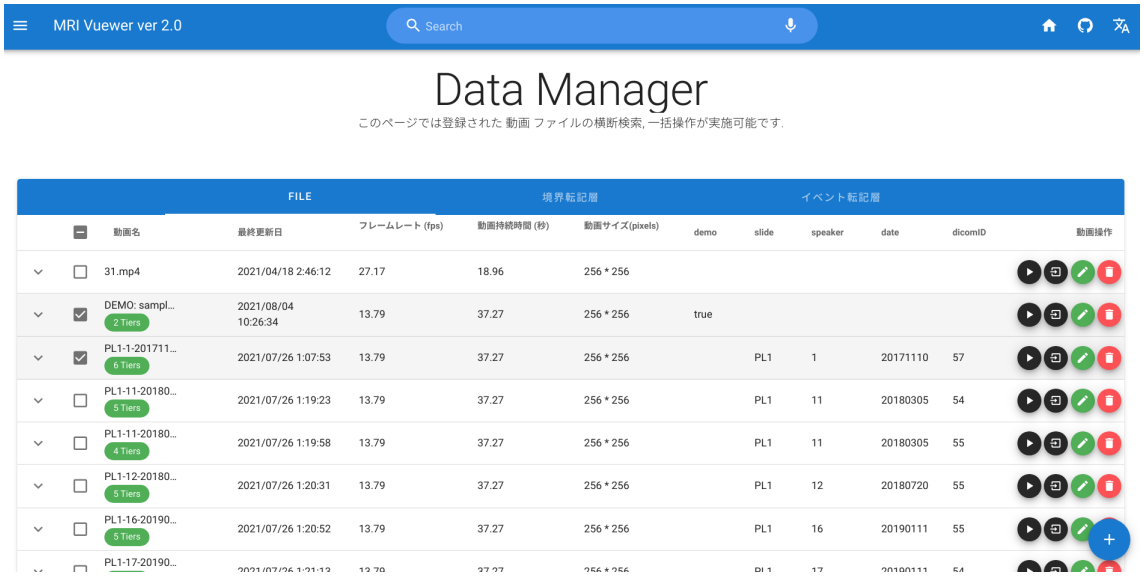


図 7　データマネージャー画面．解析者はこの画面から動画に対し横断的に検索を行ったり, 動画に対する一括処理を行うことが可能である．

が入るとアンド検索となる．検索条件を絞り込みたい場合, key=val の形式で文字列を入力する．たとえば図 7 には slide というメタ情報が登録されているが, この値が PL1 である動画のみを検索したい場合, 検索欄に slide=PL1 と入力する．

メタ情報はこのテーブルから編集可能である．テーブル右端にあるボタン群の内, 緑色の編集ボタンをクリックすると, メタ情報登録ダイアログが立ち上がる．このダイアログは大きく 2 つの入力フォームが存在する．1 つは NEW Field と書かれたフォームである．このフォームは今までに登録のないメタ情報の登録を開始するためのフォームである．入力欄に任意の文字列を入力し, ADD と書かれたボタンをクリックすると, その文字情報が “key” として登録される．もう 1 つのフォームには既存の “key” の数と一致する入力欄が表示される．ここに該当の “key” に対する “value” を登録する．最後に OK と書かれたボタンをクリックすると選択された動画データに対しメタデータが登録される．なお, この “value” は検索時には文字列として前方一致をすることに注意されたい．

加えてこのテーブルから時系列アノテーションに利用する層を一括で作成可能である．転記層を作成したい動画の行左端にあるチェックボックスを選択すると, テーブル下に転記階層一括付与と書かれた入力欄が表示される．ここに 階層名：interval—point のようなフォーマットで任意の層を指定する．なお, 複数の層を作成したい場合には上記フォーマットをスペースで区切って入力する．最後に紙飛行機のアイコンをクリックすると転記層が作成される．ただし, 既存かつ同名の転記層が存在する場合, 転記情報が初期化されることに注意が必要である．

境界転記層もしくはイベント転記層タブが活性化している場合には, 時系列アノテーションの内, それぞれの区分のデータがテーブル形式で表示される (図 8)．検索操作は前述の FILE タブが活性化している場合と同様である．テーブル右端にあるボタン群は再生ボタン, 動画切り出しボタン, アノテーション開始ボタンである．再生ボタンをクリックすると該当区間の動

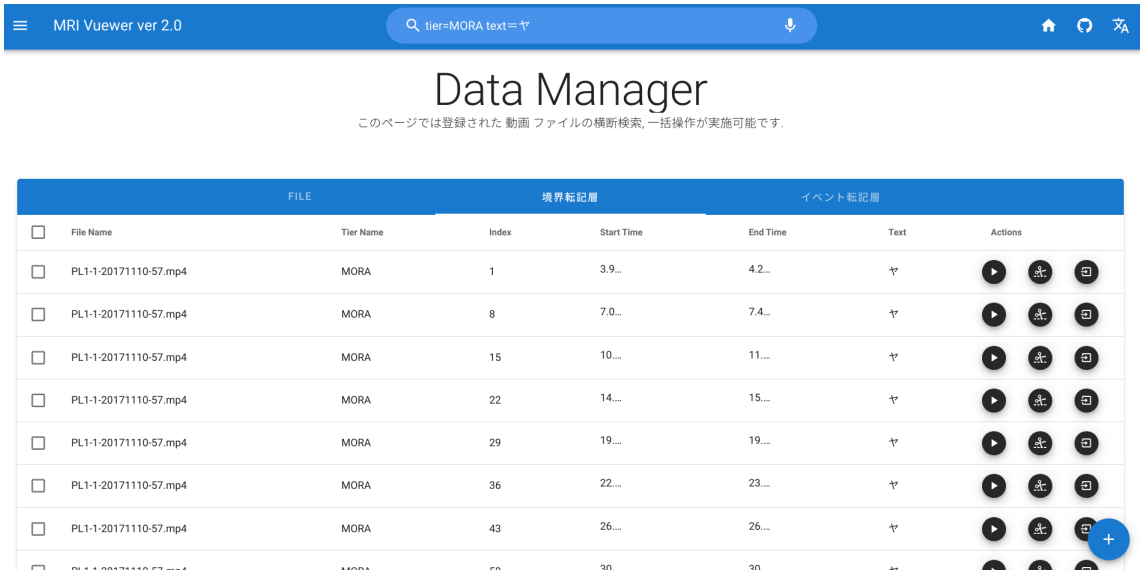


図 8　データマネージャー画面 (境界転記タブ活性化時). 境界転記タブを選択することで時間的情報の検索を実施することができる. この例では, 境界転記層が MORA の時間的情報の内, /ヤ/ と記述されたものを検索している. 解析者は左端のチェックボックスを選択することで, その区間の動画や画像をダウンロードすることができる.

画が再生される. 動画切り出しボタンをクリックすると, 該当区間の動画ファイルが切り出され, ダウンロードできる. アノテーション開始ボタンをクリックすると時系列アノテーション画面に遷移する. ただし, 動画再生コンポーネントおよび音声スペクトル表示コンポーネントの時刻情報は選択された時系列アノテーション情報の時刻情報に一致する.

テーブル左端にあるチェックボックスを選択するとテーブル下部に動画および画像のダウンロードボタンが表示される. これらのボタンをクリックすると, チェックされた動画の該当区間における, 動画や, 画像を切り出し, ZIP ファイルとしてまとめてダウンロード可能である. ただし, 動画の切り出しや画像を切り出し処理は比較的多くのマシンパワーを要求され, 解析者が使用する PC スペックや選択された行の数によっては, かなり長い時間が掛かる場合があることに注意されたい.

4.　まとめと今後の課題

本稿では, リアルタイム MRI 撮像で収録された音声付き動画データの分析環境の拡充を目標に Web アプリケーションとして機能するアノテーションツールの設計と実装を紹介した.

近年の rtMRI 研究に使用される解析手法を概括し, それらの解析手法を実施するにあたり必要になる前処理をまとめた. ここで整理された前処理, アノテーション作業の実施に際し, 必要なデータ設計を行い, 技術的な制約および実装の対する技術的背景を整理した. また, アノテーション作業をアプリケーション利用シーンとして整理し, それらの利用シーンに沿ってアノテーション実装を紹介した.

本アプリケーションは Web アプリケーションであり, 解析者は特別なツールのインストー

ルを必要とせず, 調音動画データの解析を開始することが可能になる．ここで作成したアノテーションデータは動画に対し横断的に検索可能であり, XLSX を始めとするさまざま形式のファイルとしてダウンロード可能である．言い換えれば, 本アプリケーションは解析者に, その興味に特化したデータベースの作成および検索を可能にし, ファイル単位でのデータアクセス環境を提供するツールである．

今後の課題としては画像および動画の正規化処理や, 画像データそのものの検索処理機構の拡充, そして各種アノテーション情報の自動付与のためのツール連携機能の拡充が挙げられる．

2 章で述べたとおり, rtMRI の解析を行う場合, 発話者間, あるいは特定の音声イベント発生時の画像データに対する正規化処理は重要な問題である．これらの正規化処理には規格化された普遍的な手法は確立されていない．しかし, 参照点を利用した幾何学的な変換処理 (たとえば Maekawa (2021) など) であれば, 本アプリケーションで提供しているフレームアノテーション機能や, 画像処理モジュールで利用している OpenCV, I/O モジュールで利用している FFMPEG の機能を組み合わせることで実現可能である．

画像データそのものの検索処理機構も現状の MRI Vuewer では, 改良の余地が残る．ここでいう画像データそのものの検索とは, たとえば, 単一の画像の特定範囲において輝度重心にあたる座標を検索したり, 二値化処理が実施された画像の特定範囲において白色になっている箇所の最高点や色境界の座標を検索する処理である．これらの検索処理はフレームアノテーションを簡便に実施するための補助的な機能に成り得るため, アプリケーション利用の要求に応じて, 適宜追加予定である．

アノテーション情報の自動付与は大規模な音声データベースにとって重要な問題になる．音声データに対しては Julius (Lee and Kawahara 2009), 画像データに対しては OpenCV や Dlib (King 2009), DeepLabv3 (Chen et al. 2018) のような機械学習フレームワークが存在する．本アプリケーションにおいて出力可能なファイル群を, 適切に変換すれば, 上記のような機械学習フレームワークを利用可能であるが, その変換処理を行うにはこれらの機械学習フレームワークに対する深い理解が必要になる．これは変換処理の困難さはデータマネージャー画面のファイル出力オプションの追加や, I/O モジュールの拡張によってある程度解消可能な問題であり, 適宜機能追加予定である．

謝　辞

本研究は, 日本学術振興会科学研究費 (17H02339 および 20H01265, いずれも代表者は前川) により実施しました．

文　献

Shigeru Kiritani, Kenji Itoh, and Osamu Fujimura (1975). “Tongue-pellet tracking by a computer-controlled x-ray microbeam system.” *The Journal of the Acoustical Society of America*, 57:6, pp. 1516–1520.

William J Hardcastle (1972). “The use of electropalatography in phonetic research.” *Phonetica*, 25:4, pp. 197–215.

Joseph S Perkell, Marc H Cohen, Mario A Svirsky, Melanie L Matthies, Iñaki Garabieta, and Michel TT Jackson (1992). “Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements.” *The Journal of the Acoustical Society of America*, 92:6, pp. 3078–3096.

Alan A Wrench (2000). “A Multi-Channel/Multi-Speaker Articulatory Database for Continuous Speech Recognition Research..” *Phonus*.

Maureen Stone, and Edward P Davis (1995). “A head and transducer support system for making ultrasound images of tongue/jaw movement.” *The Journal of The Acoustical Society of America*, 98:6, pp. 3107–3112.

Douglas H Whalen, Khalil Iskarous, Mark K Tiede, David J Ostry, Heike Lehnert-LeHouillier, Eric Vatikiotis-Bateson, and Donald S Hailey (2005). “The Haskins optically corrected ultrasound system (HOCUS).”.

Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, and Shrikanth S Narayanan (2013a). “An investigation of articulatory setting using real-time magnetic resonance imaging.” *The Journal of the Acoustical Society of America*, 134:1, pp. 510–519.

Vikram Ramanarayanan, Sam Tilsen, Michael Proctor, Johannes Töger, Louis Goldstein, Krishna S Nayak, and Shrikanth Narayanan (2018). “Analysis of speech production real-time MRI.” *Computer Speech & Language*, 52, pp. 1–22.

Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, and Shrikanth S Narayanan (2013b). “An investigation of articulatory setting using real-time magnetic resonance imaging.” *The Journal of the Acoustical Society of America*, 134:1, pp. 510–519.

Vikram Ramanarayanan, Maarten Van Segbroeck, and Shrikanth S Narayanan (2016). “Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories.” *Computer speech & language*, 36, pp. 330–346.

Adam C Lammert, Michael I Proctor, and Shrikanth S Narayanan (2010). “Data-driven analysis of realtime vocal tract MRI using correlated image regions.” *Eleventh Annual Conference of the International Speech Communication Association*.

Sam Tilsen, Pascal Spincemaille, Bo Xu, Peter Doerschuk, Wen-Ming Luh, Elana Feldman, and Yi Wang (2016). “Anticipatory posturing of the vocal tract reveals dissociation of speech movement plans from linguistic units.” *PloS one*, 11:1, p. e0146813.

Shinji Maeda (1979). “Un modele articulatoire de la langue avec des composantes lineaires..”10eme Journees d ’ Etude Sur la Parole, pp. 1–9.

Michael I Proctor, Daniel Bone, Athanasios Katsamanis, and Shrikanth S Narayanan (2010). “Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis.” *Eleventh Annual Conference of the International Speech Communication Association*.

Rafael De Assuncao Sampaio, and Marcel Parolin Jackowski (2017). “Vocal tract morphology using real-time magnetic resonance imaging.” *2017 30th SIBGRAPI Conference on*

Graphics, Patterns and Images (SIBGRAPI), pp. 359–366., IEEE.

Zeynab Raeesy, Sylvia Rueda, Jayaram K Udupa, and John Coleman (2013). "Automatic segmentation of vocal tract MR images." *2013 IEEE 10th International Symposium on Biomedical Imaging*, pp. 1328–1331., IEEE.

Hironori Takemoto, Tsubasa Goto, Yuya Hagihara, Sayaka Hamanaka, Tatsuya Kitamura, Yukiko Nota, and Kikuo Maekawa (2019). "Speech Organ Contour Extraction Using Real-Time MRI and Machine Learning Method.." *Interspeech*, pp. 904–908.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes (2006). "ELAN: a professional framework for multimodality research." *LREC*, pp. 1556–1559.

Paul Boersma, and David Weenink (2018). "Praat: doing phonetics by computer.".

Hanae Koiso, Yasuharu Den, Ken'ya Nishikawa, and Kikuo Maekawa (2014). "Design and development of an RDB version of the Corpus of Spontaneous Japanese.." *LREC*, pp. 1471–1476.

W3C (2021). *Indexed Database API 3.0*. `https://www.w3.org/TR/IndexedDB/`.

Guisch, and thijstriemstra (2018). *wavesurfer.js*. `https://wavesurfer-js.org`.

Suramya Tomar (2006). "Converting video formats with FFmpeg." *Linux Journal*, 2006:146, p. 10.

MDN web docs (2018a). *Promise*. `https://developer.mozilla.org/ja/docs/WebAssembly`.

G. Bradski (2000). "The OpenCV Library." *Dr. Dobb's Journal of Software Tools*.

MDN web docs (2018b). *¡video¿: The Video Embed element*. `https://developer.mozilla.org/en-US/docs/Web/HTML/Element/video`.

Kikuo Maekawa (2021). "Production of the utterance-final moraic nasal in Japanese: A real-time MRI study." *Journal of the International Phonetic Association*, pp. 1–24.

Akinobu Lee, and Tatsuya Kawahara (2009). "Recent development of open-source speech recognition engine julius." *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pp. 131–137., Asia-Pacific Signal and Information Processing Association, 2009 Annual ….

Davis E. King (2009). "Dlib-ml: A Machine Learning Toolkit." *Journal of Machine Learning Research*, 10, pp. 1755–1758.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam (2018). "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *ECCV*.

関連 URL

MRI Vuewer ver1 `https://kikuchiken-waseda.github.io/MRIVuewer/`

MRI Vuewer ver2 `https://kikuchiken-waseda.github.io/mri-vuewer.ver2/`